Classification and prediction for the breast cancer tumors

Report by Xinyi Xie, Kexin Fan, Jiawei Zeng

Background

- Cancer is a disease in which cells in the body grow out of control. Breast cancer is one of the most common cancers diagnosed. Many of us think of breast cancer as a female disease, but it can also occur in men. We filtered out some of the features in the progress of searching data. Based on these features, we can determine whether a tumor is benign(B) or malignant(M).

Data & objective

- Features are computed from a digitized image of a fine needle aspirate of a breast mass.
- Diagnosis as attribute factors and 10 real-valued features.
- No missing attribute values

- Our objective of this project is to investigate the most optimal way to classify tumors as benign and malignant, provide an understanding of the process of organizing and preparing data, selecting features and applying machine learning tools.

Data preparation

1. Clean and Generate data
2. Visualizing data

Ensemble

The key objective of the ensemble methods is to reduce bias and variance.

Ensembles of Classifiers is combining the classification results from different classifiers to produce the final output.

Four type: Stacking, Blending, Bagging and Boosting

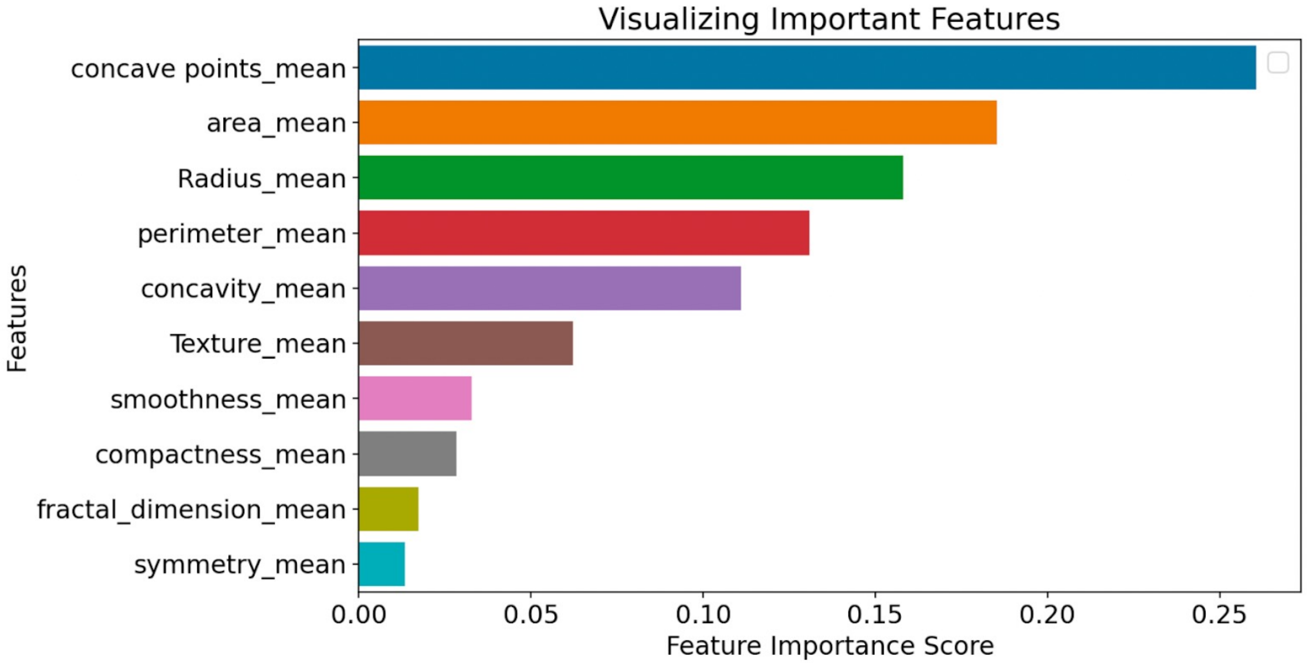
DECISION TREE(BASE MODEL)

- Accuracy score: 91.228% & Recall score is 0.97115

RANDOM FOREST(BAGGING)

- Accuracy score: 94.152% & Recall score is 0.98077

Random forest method is better than decision tree



Drop 4 variable which score less than 0.05

Bagging Algorithm:

- Sample N times from the training set
- Each sample is drawn randomly with replacement (a bootstrap).
- Learn a classifier on each bootstrap sample set
- For prediction: Make the classifiers vote

Accuracy score: 0.9532

Recall score: 0.9455

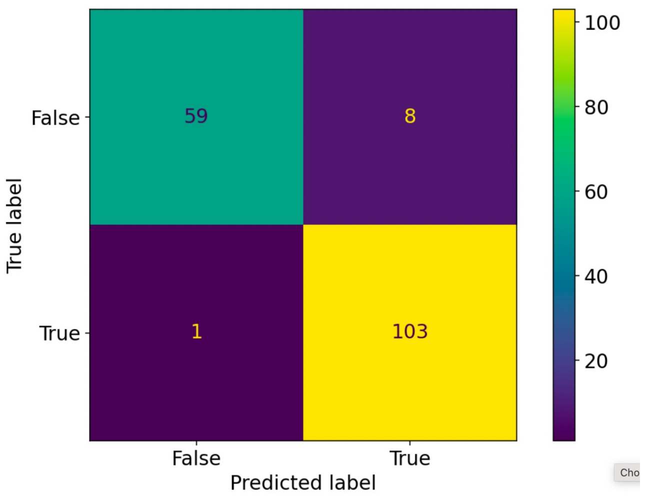
ADABOOST(BOOSTING)

Algorithm:

- Train a sequence of classifiers.
- A new classifier should **focus on those cases which were incorrectly classified in the last round**.
- Such focus is implemented by using a **weighted dataset**.

Weight:

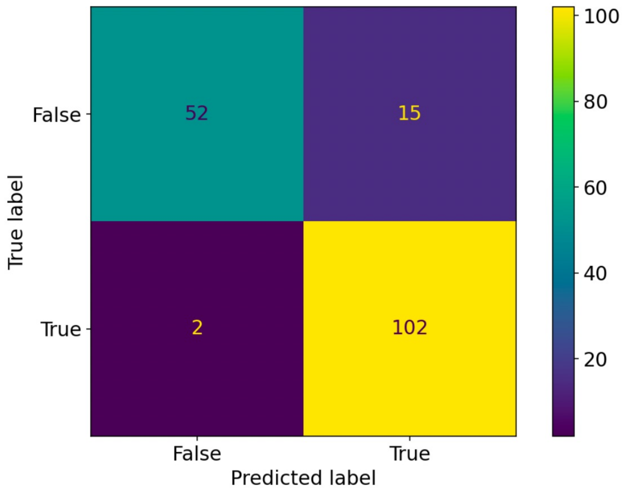
- Each instance i has an associated weight
- Model get correct, weight decrease; incorrect, weight increase.
- These weights are used as a distribution over which the dataset is sampled to create a replicated training set,



Based on decision tree :

Accuracy score:
0.94737

Recall score:
0.99034



Based on SVM(space):

Accuracy score:
0.90058

Recall score:
0.98077

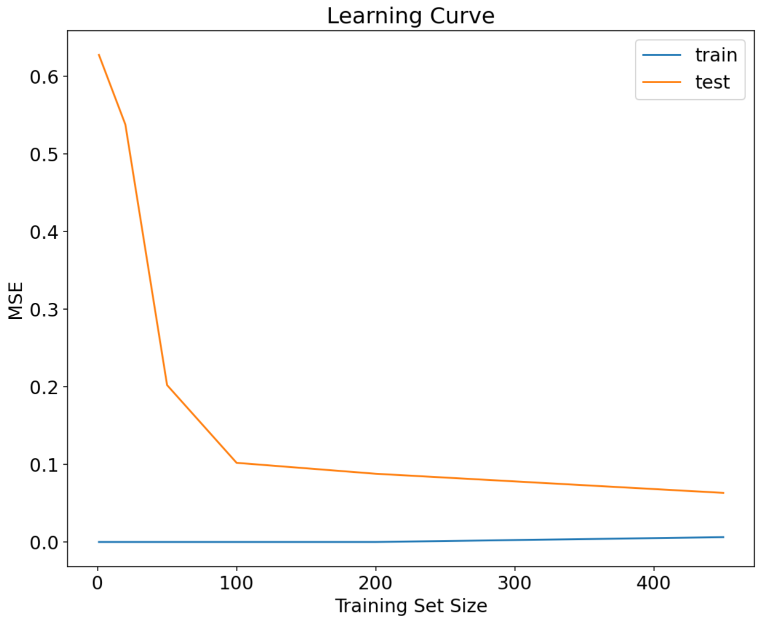
Comparison & Conclusion

Comparison:

- Tree: Parallel independently; dependent on previous
- Vote: Combine all models; some models by weight

Conclusion:

- Adaboost method with decision tree method is best for us to classify whether the tumors are benign or malignant.



Test sample: 100, MSE small